11

Interpreting test scores

11.1 Frequency distribution

Marks awarded by counting the number of correct answers on a test script are known as raw marks. '15 marks out of a total of 20' may appear a high mark to some, but in fact the statement is virtually meaningless on its own. For example, the tasks set in the test may have been extremely simple and 15 may be the lowest mark in a particular group of scores.

TABLE 1

Testee	Mark
A	20
B	25
C	33
D	35
E	29
F	25
G	30
H	26
I	19
J	27
K	26
L	32
M	34
N	27
O	27
P	29
Q	25
R	23
S	30
T	26
U	22
V	23
W	33
X	26
Y	24
Z	26

TABLE 2

Testee	Mark	Rank
D	35	1
M	34	2
C	33	3.5 (or 3=)
W	33	3.5 (or 3=)
L	32	5
G	30	6.5 (or 6=)
S	30	6.5 (or 6=)
E	29	8.5 (or 8=)
P	29	8.5 (or 8=)
J	27	11 (or 10=)
N	27	11 (or 10=)
O	27	11 (or 10=)
H	26	15 (or 13=)
K	26	15 (or 13=)
T	26	15 (or 13=)
X	26	15 (or 13=)
Z	26	15 (or 13=)
B	25	19 (or 18=)
F	25	19 (or 18=)
Q	25	19 (or 18=)
Y	24	21
R	23	22.5 (or 22=)
V	23	22.5 (or 22=)
U	22	24
A	20	25
I	19	26

TABLE 3

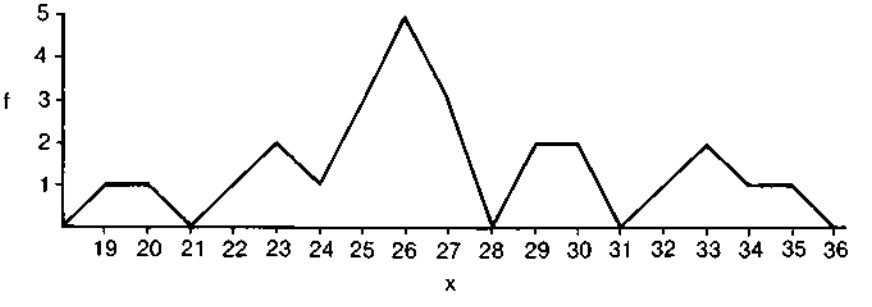
Mark	Tally	Frequency
40		
39		
38		
37		
36		
35	/	1
34	/	1
33	//	2
32	/	1
31		
30	//	2
29	//	2
28		
27	///	3
26	////	5
25	///	3
24	/	1
23	//	2
22	/	1
21		
20	/	1
19	/	1
18		
17		
16		
15	Total	26

Conversely, the test may have been extremely difficult, in which case 15 may well be a very high mark Numbers still exert a strange and powerful influence on our society, but the shibboleth that 40 per cent should always represent a pass mark is nevertheless both surprising and disturbing

The tables on the previous page contain the imaginary scores of a group of 26 students on a particular test consisting of 40 items Table 1 conveys very little, but Table 2, containing the students' scores in order of merit, shows a little more Table 3 contains a frequency distribution showing the number of students who obtained each mark awarded, the strokes on the left of the numbers (e g ////) are called *tallies* and are included simply to illustrate the method of counting the frequency of scores Note that normally the frequency list would have been compiled without the need for Tables 1 and 2, consequently, as the range of highest and lowest marks would then not be known, all the *possible* scores would be listed and a record made of the number of students obtaining each score in the scale (as shown in the example)

Note that where ties occur in Table 2, two ways of rendering these are shown The usual classroom practice is that shown in the parentheses Where statistical work is to be done on the ranks, it is essential to record the average rank (e g testees J, N and O, each with the same mark, occupy places 10, 11 and 12 in the list, averaging 11)

The following frequency polygon illustrates the distribution of the scores



11 2 Measures of central tendency

Mode

The *mode* refers to the score which most candidates obtained in this case it is 26, as five testees have scored this mark

Median

The *median* refers to the score gained by the middle candidate in the order of merit in the case of the 26 students here (as in all cases involving even numbers of testees), there can obviously be no middle person and thus the score halfway between the lowest score in the top half and the highest score in the bottom half is taken as the median The median score in this case is also 26

Mean

The *mean* score of any test is the arithmetical average i e the sum of the separate scores divided by the total number of testees The mode, median, and mean are all measures of central tendency The mean is the most efficient measure of central tendency, but it is not always appropriate

In the following Table 4 and formula, note that the symbol x is used to denote the score, N the number of the testees, and m the mean The

symbol f denotes the frequency with which a score occurs The symbol Σ means *the sum of*

TABLE 4		
x	f	fx
35	× 1	35
34	× 1	34
33	× 2	66
32	× 1	32
30	× 2	60
29	× 2	58
27	× 3	81
26	× 5	130
25	× 3	75
24	× 1	24
23	× 2	46
22	× 1	22
20	× 1	20
19	× 1	19
Total =		702 = Σfx

$$m = \frac{\Sigma fx}{N} = \frac{702}{26} = 27$$

Note that x = 702 is the total number of items which the group of 26 students got right between them Dividing by N = 26 (as the formula states), this obviously gives the average

It will be observed that in this particular case there is a fairly close correspondence between the mean (27) and the median (26) Such a close correspondence is not always common and has occurred in this case because the scores tend to cluster symmetrically around a central point

11 3 Measures of dispersion

Whereas the previous section was concerned with measures of central tendency this section is related to the range or spread of scores The mean by itself enables us to describe an individual student's score by comparing it with the average set of scores obtained by a group, but it tells us nothing at all about the highest and lowest scores and the spread of marks

Range

One simple way of measuring the spread of marks is based on the difference between the highest and lowest scores Thus, if the highest score on a 50-item test is 43 and the lowest 21, the range is from 21 to 43 i e 22 If the highest score, however, is only 39 and the lowest 29, the range is 10 (Note that in both cases, the mean may be 32) The range of the 26 scores given in Section 11 1 is 35 − 19 = *16*

Standard deviation

The standard deviation (s d) is another way of showing the spread of scores It measures the degree to which the group of scores deviates from the mean, in other words, it shows how *all* the scores are spread out and thus gives a fuller description of test scores than the range, which simply describes the gap between the highest and lowest marks and ignores the information provided by all the remaining scores Abbreviations used for the standard deviation are either s d or σ (the Greek letter sigma) or s

One simple method of calculating s.d. is shown below:

$$\text{s.d.} = \sqrt{\frac{\Sigma d^2}{N}}$$

N is the number of scores and d the deviation of each score from the mean. Thus, working from the 26 previous results, we proceed to:

1. find out the amount by which each score deviates from the mean (d);
2. square each result (d^2);
3. total all the results (Σd^2);
4. divide the total by the number of testees ($\Sigma d^2/N$); and
5. find the square root of this result ($\sqrt{\Sigma d^2/N}$).

	Score	Mean Deviation (d)	Squared (d^2)
(Step 1)	35 deviates from 27 by	8	(Step 2) 64
	34	7	49
	33	6	36
	33	6	36
	32	5	25
	30	3	9
	30	3	9
	29	2	4
	29	2	4
	27	0	0
	27	0	0
	27	0	0
	26	−1	1
	26	−1	1
	26	−1	1
	26	−1	1
	26	−1	1
	25	−2	4
	25	−2	4
	25	−2	4
	24	−3	9
	23	−4	16
	23	−4	16
	22	−5	25
	20	−7	49
	19	−8	64
	702	(Step 3)	Total = 432

(Step 4) s.d. $= \sqrt{\frac{432}{26}}$

(Step 5) s.d. $= \sqrt{16.62}$

s.d. $= 4.077$

$= \underline{\underline{4.08}}$

Note: If deviations (d) are taken from the mean, their sum (taking account of the minus sign) is zero + 42 − 42 = 0. This affords a useful check on the calculations involved here.

A standard deviation of 4.08, for example, shows a smaller spread of scores than, say, a standard deviation of 8.96. If the aim of the test is simply to determine which students have mastered a particular programme

of work or are capable of carrying out certain tasks in the target language, a standard deviation of 4.08 or any other denoting a fairly narrow spread will be quite satisfactory provided it is associated with a high average score However, if the test aims at measuring several levels of attainment and making fine distinctions within the group (as perhaps in a proficiency test), then a broad spread will be required.

Standard deviation is also useful for providing information concerning characteristics of different groups If, for example, the standard deviation on a certain test is 4.08 for one class, but 8.96 on the same test for another class, then it can be inferred that the latter class is far more heterogeneous than the former

11.4 Item analysis

Earlier careful consideration of objectives and the compilation of a table of test specifications were urged before the construction of any test was attempted What is required now is a knowledge of how far those objectives have been achieved by a particular test Unfortunately, too many teachers think that the test is finished once the raw marks have been obtained But this is far from the case, for the results obtained from objective tests can be used to provide valuable information concerning:

- the performance of the students as a group, thus (in the case of class progress tests) informing the teacher about the effectiveness of the teaching,
- the performance of individual students; and
- the performance of each of the items comprising the test.

Information concerning the performance of the students as a whole and of individual students is very important for teaching purposes, especially as many test results can show not only the types of errors most frequently made but also the actual reasons for the errors being made As shown in earlier chapters, the great merit of objective tests arises from the fact that they can provide an insight into the mental processes of the students by showing very clearly what choices have been made, thereby indicating definite lines on which remedial work can be given.

The performance of the test items, themselves, is of obvious importance in compiling future tests Since a great deal of time and effort are usually spent on the construction of good objective items, most teachers and test constructors will be desirous of either using them again without further changes or else adapting them for future use. It is thus useful to identify those items which were answered correctly by the more able students taking the test and badly by the less able students. The identification of certain difficult items in the test, together with a knowledge of the performance of the individual distractors in multiple-choice items, can prove just as valuable in its implications for teaching as for testing.

All items should be examined from the point of view of (1) their difficulty level and (2) their level of discrimination.

Item difficulty

The *index of difficulty* (or *facility value*) of an item simply shows how easy or difficult the particular item proved in the test. The index of difficulty (FV) is generally expressed as the fraction (or percentage) of the students who answered the item correctly. It is calculated by using the formula:

$$FV = \frac{R}{N}$$

R represents the number of correct answers and N the number of students taking the test. Thus, if 21 out of 26 students tested obtained the correct answer for one of the items, that item would have an index of difficulty (or a facility value) of 77 or 77 per cent

$$FV = \frac{20}{26} = \underline{77}$$

In this case, the particular item is a fairly easy one since 77 per cent of the students taking the test answered it correctly. Although an average facility value of .5 or 50 per cent may be desirable for many public achievement tests and for a few progress tests (depending on the purpose for which one is testing), the facility value of a large number of individual items will vary considerably While aiming for test items with facility values falling between 4 and 6, many test constructors may be prepared in practice to accept items with facility values between 3 and 7 Clearly, however, a very easy item, on which 90 per cent of the testees obtain the correct answer, will not distinguish between above-average students and below-average students as well as an item which only 60 per cent of the testees answer correctly On the other hand, the easy item will discriminate amongst a group of below-average students, in other words, one student with a low standard may show that he or she is better than another student with a low standard through being given the opportunity to answer an easy item Similarly, a very difficult item, though failing to discriminate among most students, will certainly separate the good student from the very good student

A further argument for including items covering a range of difficulty levels is that provided by motivation While the inclusion of difficult items may be necessary in order to motivate the good student, the inclusion of very easy items will encourage and motivate the poor student In any case, a few easy items can provide a 'lead-in' for the student – a device which may be necessary if the test is at all new or unfamiliar or if there are certain tensions surrounding the test situation

Note that it is possible for a test consisting of items each with a facility value of approximately .5 to fail to discriminate at all between the good and the poor students If, for example, half the items are answered correctly by the good students and incorrectly by the poor students while the remaining items are answered incorrectly by the good students but correctly by the poor students, then the items will work against one another and no discrimination will be possible. The chances of such an extreme situation occurring are very remote indeed; it is highly probable, however, that at least one or two items in a test will work against one another in this way

Item discrimination

The discrimination index of an item indicates the extent to which the item discriminates between the testees, separating the more able testees from the less able The index of discrimination (D) tells us whether those students who performed well on the whole test tended to do well or badly on each item in the test It is presupposed that the total score on the test is a valid measure of the student's ability (i e the good student tends to do well on the test as a whole and the poor student badly) On this basis, the score on the whole test is accepted as the criterion measure, and it thus becomes possible to separate the 'good' students from the 'bad' ones in performances on individual items If the 'good' students tend to do well on

an item (as shown by many of them doing so – a frequency measure) and the 'poor' students badly on the same item, then the item is a good one because it distinguishes the 'good' from the 'bad' in the same way as the total test score This is the argument underlying the index of discrimination

There are various methods of obtaining the index of discrimination all involve a comparison of those students who performed well on the whole test and those who performed poorly on the whole test However, while it is statistically most efficient to compare the top 27½ per cent with the bottom 27½ per cent, it is enough for most purposes to divide small samples (e g class scores on a progress test) into halves or thirds For most classroom purposes, the following procedure is recommended

1 Arrange the scripts in rank order of total score and divide into two groups of equal size (i e the top half and the bottom half) If there is an odd number of scripts, dispense with one script chosen at random
2 Count the number of those candidates in the upper group answering the first item correctly, then count the number of lower-group candidates answering the item correctly
3 Subtract the number of correct answers in the lower group from the number of correct answers in the upper group i e find the difference in the proportion passing in the upper group and the proportion passing in the lower group
4 Divide this difference by the total number of candidates in one group

$$D = \frac{\text{Correct U} - \text{Correct L}}{n}$$

(D = Discrimination index, n = Number of candidates in one group*, U = Upper half and L = Lower half The index D is thus the difference between the proportion passing the item in U and L)
5 Proceed in this manner for each item

The following item, which was taken from a test administered to 40 students, produced the results shown

I left Tokyo Friday morning
A in (B) on C at D by

$$D = \frac{15 - 6}{20} = \frac{9}{20} = \underline{.45}$$

Such an item with a discrimination index of .45 functions fairly effectively, although clearly it does not discriminate as well as an item with an index of .6 or .7 Discrimination indices can range from + 1 (= an item which discriminates perfectly – i e it shows perfect *correlation* with the testees' results on the whole test) through 0 (= an item which does not discriminate in any way at all) to −1 (= an item which discriminates in entirely the wrong way) Thus, for example, if all 20 students in the upper group answered a certain item correctly and all 20 students in the lower group got the wrong answer, the item would have an index of discrimination of 1.0

*The reader should carefully distinguish between n (= the number of candidates in either the U or L group) and N (= the number in the whole group) as used previously Obviously n = 1/2 N

If, on the other hand, only 10 students in the upper group answered it correctly and furthermore 10 students in the lower group also got correct answers, the discriminaton index would be 0 However, if none of the 20 students in the upper group got a correct answer and all the 20 students in the lower group answered it correctly, the item would have a negative discrimination, shown by −1 0 It is highly inadvisable to use again, or even to attempt to amend, any item showing negative discrimination Inspection of such an item usually shows something radically wrong with it

Again, working from actual test results, we shall now look at the performance of three items The first of the following items has a high index of discrimination, the second is a poor item with a low discrimination index, and the third example is given as an illustration of a poor item with negative discrimination

1 High discrimination index

NEARLY When Jim crossed the road, he ✓ ran into a car

$$D = \frac{18 - 3}{20} = \frac{15}{20} = .75 \qquad FV = \frac{21}{40} = \underline{0.525}$$

(The item is at the right level of difficulty and discriminates well)

2 Low discrimination index

If you the bell, the door would have been opened

A would ring C would have rung
(B) had rung D were ringing

$$D = \frac{3 - 0}{20} = \underline{.15} \qquad FV = \frac{3}{40} = \underline{.075}$$

(In this case, the item discriminates poorly because it is too difficult for everyone, both 'good' and 'bad')

3 Negative discrimination index

I don't think anybody has seen him

(A) Yes, someone has
(B) Yes, no one has
C Yes, none has
D Yes, anyone has

$$D = \frac{4 - 6}{20} = \frac{-2}{20} = -\underline{.10} \qquad FV = \frac{10}{40} = \underline{0.25}$$

(This item is too difficult and discriminates in the wrong direction)

What has gone wrong with the third item above? Even at this stage and without counting the number of candidates who chose each of the options, it is evident that the item was a trick item in other words, the item was far too 'clever' It is even conceivable that many native speakers would select option B in preference to the correct option A Items like this all too often escape the attention of the test writer until an item analysis actually focuses attention on them (This is one excellent reason for conducting an item analysis)

Note that items with a very high facility value fail to discriminate and thus generally show a low discrimination index The particular group of

students who were given the following item had obviously mastered the use of *for* and *since* following the present perfect continuous tense:

He's been living in Berlin . . . . . 1975.

$$D = \frac{19 - 19}{20} = \underline{0} \qquad FV = \frac{38}{40} = \underline{0.95}$$

(The item is extremely easy for the testees and has zero discrimination.)

Item difficulty and discrimination

Facility values and discrimination indices are usually recorded together in tabular form and calculated by similar procedures. Note again the formulae used:

$$FV = \frac{\text{Correct U + Correct L}}{2n} \qquad \left(\text{or } FV = \frac{R}{N}\right)$$

$$D = \frac{\text{Correct U} - \text{Correct L}}{n}$$

The following table, compiled from the results of the test referred to in the preceding paragraphs, shows how these measures are recorded.

Item	U	L	U+L	FV	U−L	D
1	19	19	38	.95	0	0
2	13	16	29	.73	−3	−.15
3	20	12	32	.80	8	.40
4	18	3	21	.53	15	.75
5	15	6	21	.53	9	.45
6	16	15	31	.77	1	.05
7	17	8	25	.62	9	.45
8	13	4	17	.42	9	.45
9	4	6	10	.25	−2	−.10
10	10	4	14	.35	6	.30
11	18	13	31	.78	5	.25
12	12	2	14	.35	10	.50
13	14	6	20	.50	8	.40
14	5	1	6	.15	4	.20
15	7	1	8	.20	6	.30
16	3	0	3	.08	3	.15
Etc.						

Items showing a discrimination index of below .30 are of doubtful use since they fail to discriminate effectively. Thus, on the results listed in the table above, only items 3, 4, 5, 7, 8, 10, 12, 13 and 15 could be safely used in future tests without being rewritten. However, many test writers would keep item 1 simply as a lead-in to put the students at ease.

Extended answer analysis

It will often be important to scrutinise items in greater detail, particularly in those cases where items have not performed as expected. We shall want to know not only why these items have not performed according to expectations but also why certain testees have failed to answer a particular item correctly. Such tasks are reasonably simple and straightforward to perform if the multiple-choice technique has been used in the test.

In order to carry out a full item analysis, or an extended answer analysis, a record should be made of the different options chosen by each student in the upper group and then the various options selected by the lower group.

If I were rich, I . . . . . work.

A. shan't B. won't (C.) wouldn't D. didn't

	U	L	U+L
A.	1	4	5
B.	2	5	7
C.	14	4	18
D.	3	7	10
	(20)	(20)	(40)

$$FV = \frac{U+L}{2n} = \frac{18}{40} = .45$$

$$D = \frac{U-L}{n} = \frac{10}{20} = .50$$

The item has a facility value of .45 and a discrimination index of .50 and appears to have functioned efficiently: the distractors attract the poorer students but not the better ones.

The performance of the following item with a low discrimination index is of particular interest:

Mr Watson wants to meet a friend in Singapore this year.
He . . . . . him for ten years.

A. knew B. had known C. knows (D.) has known

	U	L	U+L
A.	7	3	10
B.	4	3	7
C.	1	9	10
D.	8	5	13
	(20)	(20)	(40)

FV = .325
D = .15

While distractor C appears to be performing well, it is clear that distractors A and B are attracting the wrong candidates (i.e. the better ones). On closer scrutiny, it will be found that both of these options may be correct in certain contexts: for example, a student may envisage a situation in which Mr Watson is going to visit a friend whom he had known for ten years in England but who now lives in Singapore, e.g.

He knew him (well) for ten years (while he lived in England).

The same justification applies for option B.

The next item should have functioned efficiently but failed to do so: an examination of the testees' answers leads us to guess that possibly many had been taught to use the past perfect tense to indicate an action in the past taking place before another action in the past. Thus, while the results obtained from the previous item reflect on the item itself, the results here *probably* reflect on the teaching:

John F. Kennedy . . . . . born in 1917 and died in 1963.

A. is B. has been (C.) was D. had been

	U	L	U+L	
A.	0	2	2	
B.	0	3	3	FV = .625
C.	13	12	25	D = .05
D.	7	3	10	
	(20)	(20)	(40)	

In this case, the item might be used again with another group of students, although distractors A and B do not appear to be pulling much weight.

Distractor D in the following example is ineffective and clearly needs to be replaced by a much stronger distractor:

He complained that he . . . . . the same bad film the night before.

(A) had seen B. was seeing C. has seen D. would see

	U	L	U+L	
A.	14	8	22	
B.	4	7	11	FV = .55
C.	2	5	7	D = .30
D.	0	0	0	
	(20)	(20)	(40)	

Similarly, the level of difficulty of distractors C and D in the following item is far too low: a full item analysis suggests only too strongly that they have been added simply to complete the number of options required.

Wasn't that your father over there?

A. Yes, he was. C. Yes, was he.

(B) Yes, it was. D. Yes, was it.

	U	L	U+L	
A.	7	13	20	
B.	13	7	20	FV = .50
C.	0	0	0	D = .30
D.	0	0	0	
	(20)	(20)	(40)	

The item could be made slightly more difficult and thus improved by replacing distractor C by *Yes, he wasn't* and D by *Yes, it wasn't*. The item is still imperfect, but the difficulty level of the distractors will probably correspond more closely to the level of attainment being tested.

The purpose of obtaining test statistics is to assist interpretation of item and test results in a way which is meaningful and significant. Provided that such statistics lead the teacher or test constructor to focus once again on the *content* of the test, then item analysis is an extremely valuable exercise. Only when test constructors misapply statistics or become uncritically dominated by statistical procedures does item analysis begin to exert a harmful influence on learning and teaching. In the final analysis, the teacher should be prepared to sacrifice both reliability and discrimination to a limited extent in order to include in the test certain items which he or she regards as having a good 'educational' influence on

the students if, for example, their exclusion might lead to neglect in teaching what such items test

11.5 Moderating

The importance of moderating classroom tests as well as public examinations cannot be stressed too greatly No matter how experienced test writers are, they are usually so deeply involved in their work that they become incapable of standing back and viewing the items with any real degree of objectivity There are bound to be many blind-spots in tests, especially in the field of objective testing, where the items sometimes contain only the minimum of context

It is essential, therefore, that the test writer submits the test for moderation to a colleague or, preferably, to a number of colleagues Achievement and proficiency tests of English administered to a large test population are generally moderated by a board consisting of linguists, language teachers, a psychologist, a statistician, etc The purpose of such a board is to scrutinise as closely as possible not only each item comprising the test but also the test as a whole, so that the most appropriate and efficient measuring instrument is produced for the particular purpose at hand In these cases, moderation is also frequently concerned with the scoring of the test and with the evaluation of the test results

The class teacher does not have at his or her disposal all the facilities which the professional test writer has Indeed, it is often all too tempting for the teacher to construct a test without showing it to anyone, especially if the teacher has had previous training or experience in constructing examinations of the more traditional type Unfortunately, few teachers realise the importance of making a systematic analysis of the elements and skills they are trying to test and, instead of compiling a list of test specifications, tend to select testable points at random from coursebooks and readers Weaknesses of tests constructed in this manner are brought to light in the process of moderation Moreover, because there is generally more than one way of looking at something, it is incredibly easy (and common) to construct multiple-choice items containing more than one correct option In addition, the short contexts of many objective items encourage ambiguity, a feature which can pass by the individual unnoticed To the moderator, some items in a test may appear far too difficult or else far too easy, containing implausible distractors, others may contain unsuspected clues Only by moderation can such faults be brought to the attention of the test writer

In those cases where the teacher of English is working on his or her own in a school, assistance in moderation from a friend, a spouse, or an older student will prove beneficial It is simply impossible for any single individual to construct good test items without help from another person

11 6 Item cards and banks

As must be very clear at this stage, the construction of objective tests necessitates taking a great deal of time and trouble Although the scoring of such tests is simple and straightforward, further effort is then spent on the evaluation of each item and on improving those items which do not perform satisfactorily It seems somewhat illogical, therefore, to dispense with test items once they have appeared in a test

The best way of recording and storing items (together with any relevant information) is by means of small cards Only one item is entered on each card, on the reverse side of the card information derived from an item analysis is recorded e g the facility value (FV), the Index of

Discrimination (D), and an extended answers analysis (if carried out). After being arranged according to the element or skill which they are intended to test, the items on the separate cards are grouped according to difficulty level, the particular area tested, etc It is an easy task to arrange them for quick reference according to whatever system is desired Furthermore, the cards can be rearranged at any later date.

Although it will obviously take considerable time to build up an item bank consisting of a few hundred items, such an item bank will prove of enormous value and will save the teacher a great deal of time and trouble later. The same items can be used many times again, the order of the items (or options within each item) being changed each time. If there is concern about test security or if there is any other reason indicating the need for new items, many of the existing items can be rewritten In such cases, the same options are generally kept, but the context is changed so that one of the distractors now becomes the correct option. Multiple-choice items testing most areas of the various language elements and skills can be rewritten in this way, e g

(Grammar) I hope you ..... us your secret soon.

A. told B. will tell C. have told D. would tell

→ I wish you ..... us your secret soon.

A. told B. will tell C. have told D. would tell

(Vocabulary) Are you going to wear your best ... . for the party?

A. clothes B. clothing C. cloths D. clothings

→ What kind of ..... is your new suit made of?

A. clothes B. clothing C. cloth D. clothings

(Phoneme discrimination)

	beat	bit	beat
→	beat	beat	bit

(Listening comprehension) Student hears. Why are you going home?

Student reads:

A. At six o'clock.
B. Yes, I am.
C. To help my mother
D. By bus.

→ Student hears: How are you going to David's?

Student reads:

A. At six o'clock.
B. Yes, I am.
C. To help him.
D. By bus.

(Reading comprehension/ vocabulary) → Two-thirds of the country's (fuel, endeavour, industry, energy) comes from imported oil, while the remaining one-third comes from coal Moreover, soon the country will have its first nuclear power station

Two-thirds of the country's (fuel, endeavour, industry, power) takes the form of imported oil, while the remaining one-third is coal However, everyone in the country was made to realise the importance of coal during the recent miners' strike, when many factories were forced to close down.

Items rewritten in this way become new items, and thus it will be necessary to collect facility values and discrimination indices again

Such examples serve to show ways of making maximum use of the various types of test items which have been constructed, administered and evaluated In any case, however, the effort spent on constructing tests of English as a second or foreign language is never wasted since the insights provided into language behaviour as well as into language learning and teaching will always be invaluable in any situation connected with either teaching or testing.